(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平10－143541

(43)公開日　平成10年(1998)５月29日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/403	３４０Ａ
			15/40	３７０Ａ
			15/401	３１０Ｄ

審査請求　未請求　請求項の数９　ＯＬ　（全 19 頁）

(21)出願番号　特願平9－249100

(22)出願日　平成９年(1997)９月12日

(31)優先権主張番号　特願平8－243785

(32)優先日　平８(1996)９月13日

(33)優先権主張国　日本（ＪＰ）

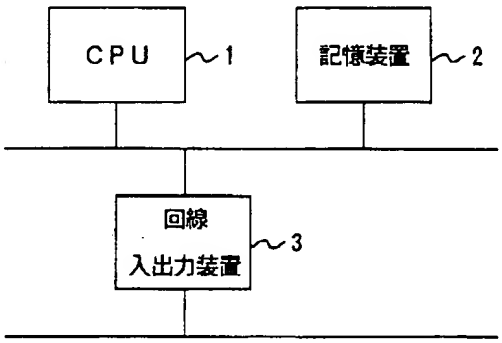
(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(72)発明者　住田　一男
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

(74)代理人　弁理士　鈴江　武彦　（外６名）

(54)【発明の名称】　情報フィルタリング装置および情報フィルタリング方法

(57)【要約】
【課題】非定期的に発生および修正される文書を対象にして、ユーザが必要とする情報のみを絞り込んでユーザに提供する情報フィルタリング装置。
【解決手段】予めプロファイル１６１に登録された検索条件と、処理対象となる文書に含まれる情報との間の類似度を類似度算出部１６が算出し、その算出した類似度にしたがって、複数の文書の中から所定の文書を選出する情報フィルタリング装置において、複数情報判定部１４は、その文書が複数の情報単位を含むか否か判定し、情報分割部１５は、複数情報判定部１４によって複数の情報単位を含むと判定された文書を情報単位ごとに分割する。そして、類似度算出部１６は、文書に対する類似度を、その文書に含まれる情報単位それぞれに算出する。したがって、複数の情報単位を含む文書内の情報単位それぞれが、回りの情報に何等影響されることなく、フィルタリング処理されることになる。



1

【特許請求の範囲】

【請求項１】　予め登録された検索条件と文書に含まれる情報との間の類似度を算出し、その算出した類似度にしたがって複数の文書の中から所定の文書を選出する情報フィルタリング装置において、

前記文書が複数の情報単位を含むか否か判定する判定手段と、

前記判定手段によって複数の情報単位を含むと判定された文書を情報単位ごとに分割する分割手段と、

前記分割手段によって分割された情報単位それぞれに、前記検索条件との間の類似度を算出する類似度算出手段とを具備してなることを特徴とする情報フィルタリング装置。

【請求項２】　複数の文書の中から所定の文書を選出する情報フィルタリング装置であって、階層構造をなすハイパーテキストをフィルタリング対象の文書に含む情報フィルタリング装置において、

新たな情報が発生したか否か監視すべき文書のアドレスを設定する第１の設定手段と、

前記第１の設定手段によって設定された文書を起点に下位層に位置する文書に対する監視すべき階層数を設定する第２の設定手段と、

前記第１の設定手段によって設定されたアドレスから前記第２の設定手段によって設定された階層数を対象範囲として文書を読み込み、その範囲内に新たな情報が発生したか否か判定する判定手段とを具備してなることを特徴とする情報フィルタリング装置。

【請求項３】　複数の文書の中から所定の文書を選出する情報フィルタリング装置において、

他の情報フィルタリング装置により出力されるフィルタリング結果を取り込む取り込み手段と、

この取り込み手段が取り込んだフィルタリング結果を前記複数の文書に含めてフィルタリング処理を実行するフィルタリング手段とを具備してなることを特徴とする情報フィルタリング装置。

【請求項４】　予め登録された検索条件と文書に含まれる情報との間の類似度を算出し、その算出した類似度にしたがって複数の文書の中から所定の文書を選出する情報フィルタリング方法において、

前記文書が複数の情報単位を含むか否か判定し、

複数の情報単位を含むと判定された文書を情報単位ごとに分割し、

この分割された情報単位それぞれに、前記検索条件との間の類似度を算出することを特徴とする情報フィルタリング方法。

【請求項５】　複数の文書の中から所定の文書を選出する情報フィルタリング方法であって、階層構造をなすハイパーテキストをフィルタリング対象の文書に含む情報フィルタリング方法において、

新たな情報が発生したか否か監視すべき文書のアドレス

2

を設定し、

この設定された文書を起点に下位層に位置する文書に対する監視すべき階層数を設定し、

前記設定されたアドレスから前記設定された階層数を対象範囲として文書を読み込み、その範囲内に新たな情報が発生したか否か判定することを特徴とする情報フィルタリング方法。

【請求項６】　複数の文書の中から所定の文書を選出する情報フィルタリング方法において、

他の情報フィルタリング装置が出力するフィルタリング結果を取り込み、

この取り込んだフィルタリング結果を前記複数の文書に含めてフィルタリング処理を実行することを特徴とする情報フィルタリング方法。

【請求項７】　予め登録された検索条件と文書に含まれる情報との間の類似度を算出し、その算出した類似度にしたがって複数の文書の中から所定の文書を選出するためのプログラムであって、

前記文書が複数の情報単位を含むか否か判定し、

複数の情報単位を含むと判定された文書を情報単位ごとに分割し、

この分割された情報単位それぞれに、前記検索条件との間の類似度を算出するようにコンピュータを動作させるプログラムを記録したコンピュータ読み込み可能な記録媒体。

【請求項８】　階層構造をなすハイパーテキストを含む複数の文書の中から所定の文書を選出するためのプログラムであって、

新たな情報が発生したか否か監視すべき文書のアドレスを設定し、

この設定された文書を起点に下位層に位置する文書に対する監視すべき階層数を設定し、

前記設定されたアドレスから前記設定された階層数を対象範囲として文書を読み込み、その範囲内に新たな情報が発生したか否か判定するようにコンピュータを動作させるプログラムを記録したコンピュータ読み込み可能な記録媒体。

【請求項９】　複数の文書の中から所定の文書を選出するためのプログラムであって、

他の情報フィルタリング装置が出力するフィルタリング結果を取り込み、

この取り込んだフィルタリング結果を前記複数の文書に含めてフィルタリング処理を実行するようにコンピュータを動作させるプログラムを記録したコンピュータ読み込み可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、膨大な数のテキスト記事や文献などの文書から、新たに入力された情報であってユーザの要求・興味にあったものを選出してユ

ーザに提供する情報フィルタリング装置および情報フィルタリング方法に関する。

【０００２】

【従来の技術】近年、インターネットの普及は目覚ましいものがあり、世界中に点在する計算機に格納された情報が、インターネットに接続されてさえいれば、どこからでも簡単にアクセスできるようになってきている。特に、WWW（Ｗｏｒｌｄ　Ｗｉｄｅ　Ｗｅｂ）では、ＨＴＴＰ（ＨｙｐｅｒＴｅｘｔ　Ｔｒａｎｓｆｅｒ　Ｐｒｏｔｏｃｏｌ）を用いることにより、利用者が、世界中の情報をＧＵＩ（Ｇｒａｐｈｉｃａｌ　Ｕｓｅｒ　Ｉｎｔｅｒｆａｃｅ）ベースのブラウザによって簡単にアクセスできる仕組みを提供している。

【０００３】WWWでは、ある計算機上でｈｔｔｐｄと呼ばれるソフトウェアを用いる。このソフトウェアは、その計算機のデータベースに格納されているＨＴＭＬ（ＨｙｐｅｒＴｅｘｔ　Ｍａｒｋｕｐ　Ｌａｎｇｕａｇｅ）で記述したハイパーテキストファイルを、他の計算機からの要求に応じて転送するものである。インターネットに接続されている計算機は、転送を要求するハイパーテキストファイルが存在するｈｔｔｐｄに対し、ハイパーテキストファイルのアドレスを指定することによって、指定したファイルを読み込むことができる。ＨＴＭＬの記述では、ハイパーテキストファイル内のリンク情報として、前記アドレスが記述されるので、ＨＴＴＰのプロトコルにしたがったブラウザは、各ｈｔｔｐｄ支配下のハイパーテキストファイルを表示することができる。そして、音声、静止画、動画などの様々なデータを出力できるようにすることによって、マルチメディアデータを含むハイパーテキストを、ブラウザは表示することができる。

【０００４】このWWWの仕組みにより、利用者は、より簡単にインターネット上の情報にアクセスできるようになり、多くの個人や企業が、Ｗｅｂページと呼ばれるハイパーテキストファイルを公開するようになってきている。

【０００５】しかしながら、WWWではデータベースの管理者がおらず、個々人がそれぞれ勝手にＷｅｂページを作成および修正し、しかもその規模が膨大であるために（１９９６年度初頭における世界中で公開されているＷｅｂページは４０００万ページと推定されている）、個々の利用者が自らが必要とするＷｅｂページがどこにあるか（ＵＲＬアドレスとして何を指定すれば必要なＷｅｂページを取得できるか）を知ることが困難な状況になっている。

【０００６】このため、最近では、アクセス可能なＷｅｂページを内容ベースで検索するシステムが開発され、検索を代行するようなサービスが行なわれるようになってきた。具体的には、Ｙａｈｏｏ、ＬｙｃｏｓおよびＡｌｔａｖｉｓｔａなどといったＷｅｂ検索サーバが存在

する。Ｗｅｂ検索サーバでは、キーワードを指定することによって、そのキーワードを含むＷｅｂページを検索することができる。利用者は、これらＷｅｂ検索サーバを用いて必要なＷｅｂページを検索する。

【０００７】しかし、このようにＷｅｂ検索サーバを用いることによってオンラインで必要な情報を容易に検索できるようになったものの、これは利用者が能動的に必要な情報を検索指示した場合にのみ得られるのであって、利用者が関心・興味をいだいている情報が新しく作成された際に利用者が検索指示を行なわなければ、たとえ重要な情報であったとしても、その利用者がその情報を知ることはない。したがって、利用者が関心・興味のある情報が発生したときに、その旨を適切な利用者に知らしめるシステムが必要である。旧来のデータベースシステムでは、このような機能をＳＤＩ（Ｓｅｌｅｃｔｉｖｅ　Ｄｉｓｓｅｍｉｎａｔｉｖｅ　Ｉｎｆｏｒｍａｔｉｏｎ）と呼んでいる。ＳＤＩでは、利用者は自らの関心・興味のある情報を選択するためのキーワードなどを個人プロファイルとしてシステムに登録しておく。そして、システムは、新しくデータが登録された際に、そのデータとキーワード（プロファイル）とを比較して、そのデータがキーワードと合致するときに、所望した情報が新たに発生した旨をプロファイルの登録利用者に知らせるものである。

【０００８】しかしながら、WWWでは、Ｗｅｂページにどのような情報を記載するかは個々人の自由であるという性格をもつために、一つのＷｅｂページに複数の情報単位が記載されることは十分に考えられる。そして、互いに関連のない複数の情報単位が記載されたＷｅｂページを一つの処理単位としてプロファイルとの比較を実行した場合、必ずしも適切なフィルタリングが施される保証はない。したがって、利用者が関心・興味をもつ極めて重要な情報が一部に含まれるＷｅｂページであっても、ページ全体としてその取捨が判定された結果、選択対象とならない場合が発生するといった問題があった。

【０００９】また、前述したような旧来型のデータベースでは、個々のデータはローカルな環境に存在するか、または特定のデータベース管理者が管理するものであったために、新しく情報が発生した情報と既存の情報とを区別することが容易であったが、WWWでは、個々人がＷｅｂページを独自に登録できる仕組みになっており、かつWWW全体を管理する管理者も存在しないため、新規情報と既存情報との区別が非常に困難である。さらに、Ｗｅｂページは、ハイパーテキスト構造をもち、互いに関連づけられた複数のページによって一定の情報を表現することがあるため、監視対象とするページについて新規情報の発生を検出するのみでは不十分であるといった問題があった。

【００１０】さらに、WWW上のＷｅｂページなどのように非常に広範囲な範囲に対して新規発生情報を監視す

ることは、単独のシステムにおいては困難であるといった問題があった。

【００１１】

【発明が解決しようとする課題】このように、従来の情報フィルタリングをたとえばWWW上のＷｅｂページなどに適用する場合においては、以下に示すような問題が存在していた。

【００１２】（１）Ｗｅｂページは単一の情報からなる場合と複数の情報からなる場合があり、複数の情報からなるページの場合に、個々の情報単位ごとに分割し、その情報単位ごとにプロファイルとの比較を行なわないと、必要な情報の選択が正確にできない。

【００１３】（２）大規模なシステムでない場合、全世界のページを網羅的にチェックすることは単独システムでは不可能である。一方、特定のページを指定して、そのページの情報が修正されたことを検出する監視手段を設けることで、利用者の便を図ることができる。しかしながら、Ｗｅｂページはハイパーテキストであるために、複数のページによって一定の情報を表現することがあり、前述の監視手段が一つのＷｅｂページだけしか指定できないと、そのページからリンクを張られている子供ページや孫ページが修正されても検出できない。

【００１４】（３）単独の情報フィルタリング装置の処理だけでは、利用者にとって十分な範囲の新規発生情報を監視することが困難である。

【００１５】この発明は、このような実情に鑑みてなされたものであり、WWWのように個々人が独自にデータを作成および修正するデータベースにおいて、新規に発生した情報（新鮮な情報）の中から、利用者の関心・興味のある情報のみを効率的に選択して通知することを可能とする情報フィルタリング装置および情報フィルタリング方法を提供することを目的とする。

【００１６】

【課題を解決するための手段】第１の発明の情報フィルタリング装置は、予め登録された検索条件と文書に含まれる情報との間の類似度を算出し、その算出した類似度にしたがって複数の文書の中から所定の文書を選出する情報フィルタリング装置において、前記文書が複数の情報単位を含むか否か判定する判定手段と、前記判定手段によって複数の情報単位を含むと判定された文書を情報単位ごとに分割する分割手段と、前記分割手段によって分割された情報単位それぞれに、前記検索条件との間の類似度を算出する類似度算出手段とを具備してなることを特徴とする。

【００１７】この第１の発明の情報フィルタリング装置においては、判定手段が、文書それぞれに対して、単一の内容からなるデータか複数の内容からなるデータかを判定する。そして、この判定手段によって複数の内容からなるデータと判定されたときに、分割手段が、その内容ごとにフィルタリング処理を行なうべく文書を情報単位ごとに分割する。そして、類似度算出手段は、この分割された情報単位それぞれに、検索条件との間の類似度を算出する。これにより、この第１の発明の情報フィルタリング装置では、単一の内容からなるＷｅｂページと複数の内容からなるＷｅｂページとに対し、これらを同時にフィルタリング対象とし、かつ内容に応じた高精度のフィルタリングを可能とすることができる。

【００１８】また、第２の発明の情報フィルタリング装置は、複数の文書の中から所定の文書を選出する情報フィルタリング装置であって、階層構造をなすハイパーテキストをフィルタリング対象の文書に含む情報フィルタリング装置において、新たな情報が発生したか否か監視すべき文書のアドレスを設定する第１の設定手段と、前記第１の設定手段によって設定された文書を起点に下位層に位置する文書に対する監視すべき階層数を設定する第２の設定手段と、前記第１の設定手段によって設定されたアドレスから前記第２の設定手段によって設定された階層数を対象範囲として文書を読み込み、その範囲内に新たな情報が発生したか否か判定する判定手段とを具備してなることを特徴とする。

【００１９】この第２の発明の情報フィルタリング装置においては、第１の設定手段が、監視すべき文書を設定し、第２の設定手段が、第１の設定手段によって設定された文書を起点とした階層数を設定する。そして、判定手段が、この第１および第２の設定手段で設定された範囲のデータを対象にフィルタリング処理を行なう。これにより、階層的なＷｅｂページを監視可能とし、指定した範囲内に新規または修正された情報があるときに、それをもれなく検知することを可能とする。

【００２０】また、第３の発明の情報フィルタリング装置は、複数の文書の中から所定の文書を選出する情報フィルタリング装置において、他の情報フィルタリング装置により出力されるフィルタリング結果を取り込む取り込み手段と、この取り込み手段が取り込んだフィルタリング結果を前記複数の文書に含めてフィルタリング処理を実行するフィルタリング手段とを具備してなることを特徴とする。

【００２１】この第３の発明の情報フィルタリング装置によれば、他の情報フィルタリング装置が出力したフィルタリング結果を取り込むことにより、単独の情報フィルタリング装置が監視できる以上の範囲の情報を監視することを可能にする。

【００２２】

【発明の実施の形態】以下、図面を参照してこの発明の実施形態について説明する。

【００２３】（第１実施形態）まず、この発明の第１の実施形態について説明する。図１に本実施形態の情報フィルタリングシステムの機器構成を示す。図１に示したように、本実施形態の情報フィルタリングシステムは、オペレーティングシステムやユーティリティを含む各種